The Pro-Truth Pledge

Fighting Fake News and Post-Truth Politics With Behavioral Science

Gleb Tsipursky[1], Fabio Votta[2]

[1]) Decision Sciences Collaborative, The Ohio State University, Columbus, OH, USA
[2]) Department of Social Sciences, University of Stuttgart, Stuttgart, Germany

We have witnessed an alarming deterioration of truth in democracies around the globe, especially in the political arena. This paper describes a proposed intervention, the Pro-Truth Pledge, which combines behavioral science research with crowd-sourcing to help address this problem. The pledge asks signers – private citizens and public figures – to commit to 12 behaviors that research in behavioral science shows correlate with an orientation toward truthfulness. Pledge mechanisms like this one have been shown in other contexts to lead private citizens to engage in more pro-social behavior. For public figures, the pledge offers specific incentives to stick to the pledge, with rewards in the form of positive reputation for honesty and truth-telling, and accountability through evaluation and potential punishment for deception. Two studies conducted on the pledge have demonstrated its effectiveness in getting private citizens to share less misinformation on social media. The pledge thus appears as an effective intervention in addressing at least some of the problems caused by fake news and post-truth politics.

Keywords: deception; post-truth politics; pro-social behavior; fake news; alternative facts

Introduction

Few would dispute that many people have lied to achieve their political agendas in the past, but this problem has become particularly bad lately. Recent political events, such as the successful tactics used by Donald Trump's campaign during the 2016 US presidential campaign and the "Vote Leave" campaign in the UK Brexit referendum, have caused the venerable Oxford Dictionary to choose as the 2016 word of the year post-truth, "circumstances in which objective facts are less influential in shaping public opinion than appeals to emotion and personal belief." (Oxford Dictionaries, 2016).

On the one hand, post-truth political methods have to do with the quantity of lies. For example, The Washington Post's well-respect Fact-Checking Column has compared the two major candidates in the US presidential election in early November 2016, and found that one of them – Trump – received their worst rating for claims fact-checking 63 percent of the time, while the other candidate – Hillary Clinton – received the worst rating 14.2 percent of the time. Notably, previously, most candidates received the worst rating between 10 and 20 percent of the time (Cillizza, 2016).

On the other, post-truth politics involves a new model of behavior when caught lying. Unlike previous politicians who backed away when caught lying, post-truth politicians do not back away from their falsehoods. Instead, they attack those who point out their deceptions, undermining public trust in credible experts and reliable news sources. This may help explain why trust among Republicans in the media has fallen by more than half, from 32 to 14 percent, from September 2015 to September 2016 (Swift, 2016).

This is not only a problem with public figures: fake news, more recently termed "viral deception" by Kathleen Hall Jamieson, director of the Annenberg Public Policy Center, is sweeping social media, shared by ordinary citizens (The Annenberg Public Policy Center, 2017). Sharing such misinformation – at least by private citizens – is not necessarily intended to harm others or even deliberately deceive. Our emotions and intuitions focus more on protecting our worldview and personal identity, and less on finding out the most accurate information (Haidt, 2012; McDermott, 2004; Nyhan & Reifler, 2010). Still, regardless of the intentions, the impact of sharing this misinformation is vast. A study showed that in the three months before the US presidential election, the top 20 fake election-related news stories on Facebook received more engagements – reactions, comments, and shares – than the top 20 real news stories, 8,711,000 compared to 7,367,000 (Silverman, 2016). Thus, real news in this case has been outcompeted by fake news. Another study, which looked at a wider number of fake news stories, showed that

in the same period of three months before the 2016 US Presidential election, 156 misleading news stories got just under 38 million shares on Facebook (Allcott & Gentzkow, 2017). Note that the researchers in this study examined shares on Facebook rather than engagements: the latter number would have been much higher. Fake news tends to favor conservative perspectives: the same study showed that stories favorable to Donald Trump were shared 30 million times, while those favorable to Hillary Clinton were shared a total of 7.6 million times. The specific impact of candidates on their supporters sharing false information may be explained in part by the research on emotional contagion, which shows that followers tend to absorb and emulate the emotions of their leaders (Hatfield, Cacioppo, & Rapson, 1993).

How impactful is such sharing? First, we need to recognize that most US adults get news on social media, 62 percent according to a recent poll by the Pew Research Center (Gottfried & Shearer, 2016). Another recent poll by Ipsos, conducted in late November and early December 2016, showed that American adults are prone to be deceived by fake news headlines. Surveying 3,015 adults, the method of this poll involved showing respondents six election-related headlines, three false and three true, and asked if they recognized the headline. In the case that they did recognize the headline, the respondents were asked to rate the headline as "very accurate," "somewhat accurate," "not very accurate," or "not at all accurate." Of the people who recognized the fake election-related headlines, approximately 75 percent rated the headlines "very accurate" or "somewhat accurate." Republicans were slightly more likely to be fooled by fake news, rating fake news headlines they recalled as accurate 84 percent of the time, compared to 71 percent for Democrats (Silverman & Singer-Vine, 2016).

This fake news came from a variety of channels, but a major portion came from Russia's efforts to use its digital propaganda efforts to influence the US election. US intelligence agencies uniformly agree about Russia's role. Recent US congressional investigations also shed light on Russia's successful efforts (Kwong, 2017). Of course, political partisans for either side, but especially Republicans, created massive amounts of fake news (Green & Issenberg, 2016). So did people trying to make money off spreading fake news (Subramanian, 2017).

Of course, the US is far from unique in the impact of fake news. The UK was another target of Russian's digital propaganda efforts, with researchers at the University of Edinburgh finding many hundreds of accounts operating by the Russian Internet Research Agency trying to spread fake news to influence UK politics (Booth, Weaver, Hern, & Walker, 2017). Russia-owned accounts spread misinformation in Spain to stir up the Catalonian independence movement (Palmer, 2017). Russia likewise used misinformation to try to influence the German 2017 elections

(Shuster, 2017). The 2017 French elections also drew a great deal of fake news, with a substantial amount coming from Russian-backed accounts (Farand, 2017). Those outside the US are similarly highly susceptible to believing fake news when exposed to it. For example, a research study on misinformation in the 2017 French election found that exposing voting-age French people to deceptive election-related statements resulted in the study subjects believing Le Pen's falsehoods. Fact-checking improved the likelihood that people believed in the actual facts (Barrera, Guriev, Henry, & Zhuravskaya, 2017).

Since it is actively harmful for our global society for people to believe in and spread falsehoods, how can we stop this problem? A recent research article by prominent scholars in the field suggested any effort to address the situation "must involve technological solutions incorporating psychological principles, an interdisciplinary approach that we describe as 'technocognition.'" (Swire, Berinsky, Lewandowsky, & Ecker, 2017). In the meantime, a separate group of psychologists have been thinking along the same lines, and have come up with a proposed technocognitive intervention we term the Pro-Truth Pledge (PTP). The pledge asks signees to commit to 12 behaviors that research in psychology shows correlate with an orientation toward truthfulness. Early results show both that private citizens and public figures are willing to take the pledge, and interviews, external observations, and quantitative studies show evidence of the effectiveness of the pledge.

Truth and the Tragedy of the Commons

Although our society as a whole loses when deception is rampant in the public sphere, individuals who practice deceptive behaviors often gain for their own agendas. This type of situation is known as a "tragedy of the commons," following a famous article in Science by Harding (1968). Harding demonstrated that in areas where a group of people share a common resource without any controls on the use of this resource, each individual may well have a strong interest in taking more of the common resource than is their fair share, leading to individual gain at great cost to the community as a whole. A well-known tragedy of the commons is environmental pollution (Vogler, 2000). We all gain from clean air and water, yet individual polluters, from a game-theoretical perspective, may well gain more – at least in the short and medium term – from polluting our environment (Hanley & Folmer, 1998). Pollution of truth is arguably similarly devastating to the atmosphere of trust in our political environment.

Solving tragedies of the commons requires, according to Hardin, "mutual coercion, mutually agreed upon by the majority of the people affected," so as to prevent these harmful outcomes where a few gain at the cost of everyone else (Harding, 1968). The environmental movement

presents many examples of successful efforts to addressing the tragedy of the commons in environmental pollution (Ostrom, 2015). Only substantial disincentives for polluting outweigh the benefits of polluting from a game theoretical perspective (Fang-yuan 2007). Particularly illuminating is a theoretical piece by Mark van Vugt describing the application of psychology research to the tragedy of the commons in the environment. His analysis showed that in addition to mutual coercion by an external party such as the government, the commons can be maintained through a combination of providing credible information, appealing to people's identities, setting up new or changing existing institutions, and shifting the incentives for participants (Van Vugt, 2009).

The research on successful strategies used by the environmental movement fits well with work on choice architecture and libertarian paternalism. "Libertarian paternalism" refers to an approach to private and public institutions that aims to use findings from psychology about problematic human thinking patterns – cognitive biases – to shape human behavior for social good while also respecting individual freedom of choice (Sunstein & Thaler, 2003a, 2003b, 2008). Choice architecture is the method of choice used by libertarian paternalists, through shaping human choices for the welfare of society as a whole, by setting up default options, anticipating errors, giving clear feedback, creating appropriate incentives, and so on (E. Johnson et al., 2012; Jolls, Sunstein, & Thaler, 1998; Selinger & Whyte, 2011; Sunstein, Thaler, & Balz, 2010).

A Proposed Intervention to Address Pollution of Truth: The Pro-Truth Pledge

The Pro-Truth Pledge (PTP), created by a team of behavioral scientists coming primarily from a psychology background, is informed by strategies that have proven successful in the environmental movement and combines them with choice architecture. The pledge is not a way for pledge organizers to tell people what is the truth, but to get them to adopt research-informed methods meant to orient toward accurate evaluation of reality. In taking the pledge, signees agree to abide by twelve behaviors, which are intended to counteract a number of cognitive biases that contribute to people believing in and sharing misinformation, an essential aspect of the psychology research informing the content of the pledge itself. The full pledge reads as follows:

I Pledge My Earnest Efforts To:

Share the truth

- Verify: fact-check information to confirm it is true before accepting and sharing it
- Balance: share the whole truth, even if some aspects do not support my opinion

- Cite: share my sources so that others can verify my information
- Clarify: distinguish between my opinion and the facts

Honor the truth

- Acknowledge: acknowledge when others share true information, even when we disagree otherwise
- Reevaluate: reevaluate if my information is challenged, retract it if I cannot verify it
- Defend: defend others when they come under attack for sharing true information, even when we disagree otherwise
- Align: align my opinions and my actions with true information

Encourage the truth

- Fix: ask people to retract information that reliable sources have disproved even if they are my allies
- Educate: compassionately inform those around me to stop using unreliable sources even if these sources support my opinion
- Defer: recognize the opinions of experts as more likely to be accurate when the facts are disputed
- Celebrate: celebrate those who retract incorrect statements and update their beliefs toward the truth

One of the biases that the pledge aims to address is the confirmation bias, our tendency to search for and accept information that aligns with our current beliefs (Nickerson 1998). Research shows that one way to address the confirmation bias involves asking people to consider and search for evidence that disproves their initial beliefs, so that they would not violate the pledge by sharing misinformation (Hirt and Markman 1995, Kray and Galinsky 2003, Lilienfeld, Ammirati and Landfield 2009).

To ensure full clarity on what constitutes violations of the pledge, the pledge spells out what misinformation means from the perspective of the PTP: anything that goes against the truth of reality, such as directly lying, lying by omission, or misrepresenting the truth to suit one's own purposes. While sometimes misinformation is blatant, sometimes it is harder to tell, and for these tough calls, the PTP relies on credible fact-checking organizations – the same ones that Facebook uses for its fact-checking program – and/or the scientific consensus, as recognized by meta-analysis studies and statements from influential scientific organizations.[1] The pledge asks

[1]The full text of the statement on what the pledge consider misinformation is given in the FAQ of the pledge, to avoid misconceptions and provide clarity. It can be seen in the link below: Pro-Truth Pledge (n.d.). Pro-Truth

people to take time to verify information before sharing it, by going to reliable fact-checking websites or evaluating the scientific consensus on any given topic. By taking time to verify this information, signees get an opportunity to evaluate the accuracy of their information and change their perspective if they do not find credible evidence supporting that information. This aspect of the pledge aims to address the extensive sharing of fake news, both by private citizens and by public figures (Allcott & Gentzkow, 2017). It also aims to address the repeated sharing of incorrect information, which produces the illusory truth effect, people's belief that a false statement is true due to multiple exposure to and thus growing comfort with the false statement (Fazio, Brashier, Payne, & Marsh, 2015). Likewise, asking people to pause and verify before sharing information will slow down their responses, which has been correlated with making fewer errors and facilitating analytical thinking to counteract belief bias (Pennycook, Cheyne, Koehler, & Fugelsang, 2013).

In the spirit of anticipating errors, an important aspect of choice architecture, the pledge encourages signees to celebrate both others and themselves for retracting incorrect statements and updating their beliefs toward the truth. We anticipate that another problematic factor might be the in-group bias, which causes people to favor those who they perceive to be part of their own group, and vice versa for those who they perceive as part of their out-group (Mullen, Brown, & Smith, 1992; Verkuyten & Nekuee, 1999). To address the in-group bias, the pledge asks people to defend other people who come under attack for sharing accurate information even if they have different values, and to request that those who share inaccurate information retract it, even if they are their friends and allies. The Dunning-Kruger effect is another cognitive bias where those who have less expertise and skills in any given area have an inflated perception of their abilities, in other words are ignorant of their own ignorance (Dunning, 2011; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999; Sheldon, Dunning, & Ames, 2014). To address this problem, the pledge calls on signees to "recognize the opinions of those who have substantially more expertise on a topic than myself as more likely to be accurate in their assessments."

In addition to the cognitive biases that facilitate deception, other studies have emerged on motivators for honesty and dishonesty. If people perceived others around them as behaving dishonestly, they were also more likely to behave dishonestly themselves; in turn, if they behaved dishonestly, they perceived others as more likely to behave dishonestly (Gino, Norton, & Ariely, 2010). These two patterns together, once they start, create a self-reinforcing spiral of deception. For our purposes, the parallel is clear. For instance, consider social media sharing of viral

Pledge FAQs. Accessed August 28, 2017 https://www.protruthpledge.org/misinformation/

deception. A person who spreads such deceptive content will perceive others around them as more likely to spread viral deception than is actually the case; likewise, if that person sees someone sharing misinformation, they will be more likely to share viral deception themselves, as that person's actions provide him with an implicit permission to do so. Similarly applicable to spreading misinformation online, research shows that people are more likely to lie if they believe it benefits their in-group (Mazar, Amir, & Ariely, 2006). So if someone sees an article favorable to their political in-group, they would be more likely to share it without doing any fact-checking, even if the article inspires some skepticism, by comparison to a neutral article. Doing such promotion of questionable content favorable to one's in-group both helps people feel like activists for their cause, and signals to others in their social media network an alliance around shared values, gaining them social capital. Moreover, research shows that if an expert corrects erroneous information, people tend to not accept and internalize the correction unless they also trust the expert; in turn, trust alone is enough to sway people to accept corrective information (Guillory & Geraci, 2013). Thus, any proposed solution needs to address the perception of dishonesty by others and oneself, address benefits to one's in-group from dishonesty, and perceptions of trustworthiness.

Fortunately, we also have research on what causes people to avoid dishonest behavior. Two articles show some intriguing findings: reminders about ethical behavior made people less likely to lie; getting people to sign an honor code or other commitment contract to honesty before engaging in tasks involving temptation to lie increased honesty; making standards for truthful behavior clear decreased deception (mazar2008a; Mazar, Amir, & Ariely, 2008). Evidence shows that personal experience with an issue such as global climate change helps correct misinformation about it (Myers, Maibach, Roser-Renouf, Akerlof, & Leiserowitz, 2013). Reminders about the reputation costs of making false statements proved effective for reducing misinformation shared by candidates for office (Nyhan & Reifler, 2010). This finding is particularly salient to the potential impact of the pledge on public figures who have a reputation that might be negatively impacted if they are found to share misinformation, due to the accountability mechanism of pledge-takers monitoring each other, especially public figures who have taken the pledge.

In an interesting parallel to the environmental movement, those who chose to commit to recycling by signing a pledge were more likely to follow their commitments in comparison to those who just agreed to recycle (Katzev & Pardini, 1987). Our likelihood of lying is strongly impacted by our social network, making it especially important to address social norms around deception (Mann, Garcia-Rada, Houser, & D, 2014). Dan Ariely summarizes and synthesizes the research on what moves us to lie and vice versa in his The Honest Truth About Dishonesty: How We Lie

to Everyone—Especially Ourselves. In a nutshell, he finds that what determines whether people lie or not is not some rational cost-benefit analysis, but a wide variety of seemingly-irrational psychological factors. Crucially, our behavior around deception ties strongly to self-identity and group belonging. People generally wish to maintain a self-identity as essentially truthful and to act within accepted group norms, and so inducing a greater orientation toward the truth requires integrating truth-oriented behaviors into one's identity and group affiliation (Ariely & Jones, 2012). The more of these factors a solution can address, the better.

The Pro-Truth Pledge: Private Citizens

We separate the targets for the pledge signees into two categories, private citizens and public figures, and will talk about the former first. Why would private citizens take the pledge? Many people are frustrated and disheartened by the prevalence of deception in our society, and especially in our political system. Signing the pledge gives them an opportunity to express their discontent and help move our society toward greater honesty. This type of pro-social desire has been found to be a strong motivator in environmental efforts (Van Lange, De Bruin, Otten, & Joireman, 1997; Van Vugt & Samuelson, 1999). Furthermore, signing the pledge gives any individual who signs it greater credibility among their peers who know they signed it. The pledge encourages individuals who signed it to share about it on their social media and personal networks, and also put a badge on their online presence indicating they signed it (Pro-Truth Pledge Badge). They get access to unique resources available to signees, such as a search engine composed of credible sources verified as reliable by the PTP organizers (Pro-Truth Pledge Search Engine). They also get to join a variety of closed communities both online and in their local area available only to pledge signees, where they can rely on the credibility of the information being shared by those who signed the pledge and also support and encourage each other in practicing behaviors advocated by the pledge. We know that peer support has proven helpful in maintaining desired behavior change in contexts such as health behaviors, and we anticipate that such support will help maintain truth-oriented behavior (Westman, Eden, & Shirom, 1985; Zimmerman & Connor, 1989). The pledge appeals to people's identities by asking for those who self-identify as truthful and honest to take the pledge and join the community of pledge-takers. This appeal to identity is informed by psychology research on the environmental movement showed that people who report self-identification with a community tend to engage in behaviors condoned by that community (Van Vugt, 2001).

However, would pledge-takers who are private citizens, and thus have no external monitoring, follow such behaviors upon taking the pledge? Psychology research on precommitment suggests

that those who commit to a certain behavioral norm will be more likely to follow it (Ariely & Wertenbroch, 2002). Another factor at play is post-factum justification or choice-supportive bias, where our minds want to perceive our past decisions in a positive light, making us more likely to stick to past commitments (Correia & Festinger, 2014). A related phenomena is a preference for consistency, which recent research suggests influences many people to make decisions that are consistent with their past decisions (Guadagno & Cialdini, 2010). Most relevantly for the PTP, at schools that have honor codes students tend to engage in less academic dishonesty (McCabe & Trevino, 1993; McCabe, Trevino, & Butterfield, 1999). Likewise, signing an honor code before a test tends to decrease cheating compared to signing an honor code at the end of a test (L. Shu, Mazar, Gino, Ariely, & Bazerman, 2012). This evidence is further supported by research from the environmental movement on recycling, which shows that those who chose to commit to recycling by signing a pledge were likely to follow on their commitments in comparison to those who just agreed to recycle (Katzev and Pardini 1987). By analogy, we hypothesize that taking the PTP will decrease sharing misinformation by shifting the underlying mental habits of thought and feeling that contribute to deceptive behaviors, especially since we are concerned with people not sharing misinformation after they sign the pledge rather than before it (Ariely & Jones, 2012; Frijda, Manstead, & Bem, 2010).

Further strengthening precommitment, post-factum justification, and preference for consistency, pledge-takers have an opportunity to participate in PTP community-oriented activities described above, to sign up for email updates, to have themselves listed in a public database of people who signed the pledge, and to share publicly about taking the pledge. They can also sign up to be a PTP advocate, which consists of any of the following: 1) Promoting the PTP pledge to other private citizens; 2) Advocating for public figures to take the pledge; 3) Monitoring and evaluating whether the public figures stick to their commitment. In the initial sign-ups, about 85 percent signed up for email updates or action alerts, about 50 percent wanted to be listed in a public database, and about 30 percent indicated an interest in being a PTP advocate (we do not have sufficient data on community engagement).

We hypothesize that each of the four distinct activities listed above would make it more likely for people to abide by the tenets of the PTP, based on research from successful environmental movement strategies. We suspect that for those who sign the PTP without signing up for email notifications or other forms of active engagement will have a small or perhaps negligible long-term impact on their behaviors, due to the PTP fading from their mind. After all, research on health behaviors shows that intentions to change behavior often fail before temptations or lack of energy, which in the PTP context we can compare to failing to fact-check an article before sharing it

(Schwarzer, 2008). Still, given that people who have committed to recycling by signing a pledge did practice recycling at a higher rate than those who did not, we may indeed witness some impact. Other research on recycling shows that having information about conservation made people more likely to engage in recycling (Oskamp et al., 1991). Getting email updates about the PTP would serve that function. Studies on recycling also show that getting specific recycling opportunities increased the likelihood of recycling, and the action alerts fill that function for the PTP (Vining & Ebreo, 1992). Knowing that one is being monitored for recycling and may get negative messages if one does not recycle has been shown to increase recycling behavior (Lord, 1994). The parallel for the PTP is choosing to list oneself in a public PTP database and thus make oneself available for monitoring, as well as sharing with one's social network and on social media that one took the PTP. Also supportive of the importance of the latter, studies of consumers buying environmentally-friendly products showed that such purchases stemmed in part from the opportunity to signal environmental friendliness to others as a form of status-seeking, and thus sharing about the PTP would similarly signal truth-friendliness (Griskevicius, Tybur, & Bergh, 2010). Active volunteering and community engagement in recycling programs, such as block-leader programs, proved even more effective in increasing recycling behavior (Burn, 1991; Hopper & Nielsen, 1991). By analogy, we anticipate that those who engage actively in PTP volunteering and community-oriented activities, online and in-person, will be even more likely to exhibit truth-oriented behaviors. After all, community belonging is crucial for shaping perceptions of self-identity and social norms, which research has found are so important in determining truth-telling behavior.

The Pro-Truth Pledge: Public Figures

Why should public figures take the PTP? We anticipate that some public figures would be motivated by the same intrinsic motivations that would lead private citizens to take the pledge. However, we wanted to provide particular incentives for public figures to take the pledge, and also disincentives for breaking the pledge, and we decided to do so in the form of reputation. Reputational rewards and penalties have been shown to be vital in addressing tragedies of the commons in the environmental movement (Milinski, Semmann, & Krambeck, 2002). Other research also demonstrated the social benefits of coordinated punishments to sustain cooperation and prevent defection (Boyd, Gintis, & Bowles, 2010). The PTP borrows from this approach.

How are public figures rewarded for taking the pledge? Taking the pledge is a way of providing credible information about the honesty of a public figure to an audience interested in such information, thus providing a substantial reputational reward. When signing the pledge, each

public figure has an opportunity to provide a brief statement about why they took the pledge, and some links to their online presence. This information will be stored in a publicly-accessible database that anyone can access, such as constituents interested in evaluating political candidates for office or deciding whether to trust the commentary of a media figure, policy expert, or academic commenting on public affairs. Moreover, the statement would get sent in a regular newsletter to all pledge signers who chose to subscribe to email updates. Doing so improves that public figure's reputation and gains them new supporters. The public figure can provide additional content for the PTP newsletter about how the pledge changed their behavior, further reinforcing both their reputation and providing proof for the PTP newsletter subscribers of the effectiveness of the pledge, creating a virtuous cycle characteristic of successful innovations (Casadesus-Masanell & Ricart, 2011).

Such provision of information has been crucial in successful interventions within the environmental movement to address tragedies of the commons. As an example, research shows that labels on household appliances that list comparisons of energy use and emissions most effectively change behavior when consumers are already concerned with the environment but lack technical knowledge about the appliances (Dietz, Ostrom, & Stern, 2003). Similarly, many consumers of political information lack knowledge about which officials and media figures and analysts are credible, and the PTP pledge provides that information.

Many may worry about the problem of false signaling or cheating – a public figure may take the pledge to signal a commitment to the truth, without actually abiding by the pledge (Connelly, Certo, Ireland, & Reutzel, 2010). Private citizens have little incentive to take their time and share their personal data by filling out the pledge, making it likely that only those committed to advancing the cause of truth in our society would take this action. However, the reputational value for public figures of taking the pledge, especially as the PTP gains popularity and credibility and also has a bigger email list, will grow higher and higher. If we do not prevent false signaling and cheating on the pledge, the pledge will not be able to provide credible information and thus fail to shift incentives to favor sharing accurate information instead of deception.

To address cheating, the pledge involves a monitoring mechanism that makes sure the pledge has teeth in the form of reputational penalties which are commensurate with the infraction. Some PTP advocates are assigned the duty of monitoring public figures. If an advocate suspects that a public figure violated the pledge, the advocate will contact the individual privately, with an approach of "innocent until reasonably shown guilty" perspective – perhaps the person misspoke, or the advocate misunderstood. If the public figure withdraws the statement, or the advocate

finds no likely violation of the pledge, the matter ends there.

If the advocate still thinks there might be a violation of the pledge, the advocate will then escalate the matter to PTP mediating committee, depending on the stature of the public figure. While anyone who signs up to the PTP may become an advocate, mediating committees are composed of a group of vetted volunteers who will evaluate the evidence provided by the advocate, contact the public figure for a chance to offer a defense, and make a ruling. If there is a ruling of a violation, then this ruling is evaluated by a member of the PTP Central Coordinating Committee, to ensure fairness and accuracy, and provide an external perspective. In the case that the PTP Central Coordinating Committee member also determines that a violation has occurred, the committee will then contact the public figure, offering the person a final chance to retract the statement. If the public figure still refuses to take their words back, the PTP mediating committee will then consider that the public figure has made a deliberate decision to lie, and will rule the public figure to be in contempt of the pledge.

This process might sound a little convoluted, but it minimizes the possibility of the PTP being politicized or corrupted at a local level, a concern raised by many in the formulation of the pledge. Indeed, research on the environmental movement showed that for an institution such as the PTP to succeed in gaining trust and credibility, it needs to demonstrate transparent, clear, and fair rules and procedures where all participants have a chance to make their case and feel heard. For instance, research on the California water shortage in 1991 showed that people cooperated with drastic water-saving measures by local water authorities only if they believed these authorities to listen to the concerns of all and provide clear, accurate, and unbiased information (Tyler & Degoey, 1995).

Once someone is found to be in contempt of the pledge, the mediating committee will then proceed to put reputational pressure on the individual to get that individual to change their position on the matter. It would issue a press advisory to all relevant media – for instance, all the media in the San Francisco area if the public figure is the mayor of San Francisco – that the public figure is in contempt of the pledge. It will also issue an action alert to those who indicated they want to receive such alerts – either at the local, regional, or national level, depending on the stature of the public figure – for them to email, tweet, call, write, and protest in front of the office of the public figure encouraging the person to revise the relevant statement, and writing letters-to-the-editor about the situation. Finally, the public figure will be listed on the PTP website as in contempt of the pledge. We anticipate that these consequences will provide considerable reputation pressure for a public figure to avoid being in contempt of the pledge. If

the public figure envisions violating the pledge deliberately, they would be better off not signing it at all. Thus, the pledge is not simply cheap talk, as it has strong reputational pressure behind it.

So why should elected or appointed officials take the pledge if it restrains their activities and causes them to make such statements retracting their posts? Officials need to be perceived as trustworthy by citizens. The PTP provides that credibility, due to the presence of the monitoring mechanism. Citizens can easily look them up in the PTP database. If the official has signed the pledge a while ago and is not in contempt, the citizen can assume the official has not made any deceptive statements without retracting them later. The official gets additional benefits because when the official signs up, her information is included in the PTP updates. This provides the official with positive reputation as being honest and credible, and gets them more support. There is an additional benefit for elected officials whose opponent for office has not taken the PTP, since the official can raise questions about why the opponent does not wish to take the pledge. The PTP thus offers a first mover advantage for those public officials who take it early onward (Kerin, Varadarajan, & Peterson, 1992).

Politicians are already taking the pledge, with over 20 having done so already, showing its promise as a tool to shift incentives. For instance, here is a statement from one such politician, the Democrat Dan Epstein:

> As a progressive who has always valued learning to make our society better, as a Democrat who believes in ethics and transparency in government and politics, as a lifelong student and teacher who has always been devoted to the sciences, humanities, and all forms of study, I will tell the truth, promote the truth, and live the truth. I will stand against not only my opponents, but my own co-partisans if need be, to honor the truth in the face of falsehood. I am running for the US House of Representatives in the Texas 19th Congressional District in 2018. http://www.danepsteinforwesttexas.com/

Here is another one, from Republican Jay Baumeister:

> I feel it is time to bring the country back together and this can not be done the way congress is acting now in an us vs them mentality. Most congressmen have only one goal and that is to get reelected. Congressmen will say whatever they need to in order to accomplish that goal truth or not. I pledge to work toward the truth and to be willing to speak the truth even if it is not in my best interest politically. I am a Republican running for Congress in Ohio's District 12: https://www.facebook.com/Baumeister-for-Congress-1682557778660008/

As politicians, the two candidates know how to speak to different audiences. The statements are specifically crafted to appeal to people who care about honesty. So we already see the reputational incentives of the pledge working out for these two candidates.

What about policy experts, commentators, analysts, media figures, and scholars? They all need to be perceived as trustworthy by the audiences to which they communicate. The PTP provides them with that benefit due to the monitoring mechanism, and similarly to the officials described above, the longer they are signed up without being in contempt, the more credibility they get. Those who sign can also get a broader audience engaged with them since their information will be included in the PTP updates. Moreover, if their competitors do not sign the pledge, those who signed up will get a bigger audience, since audiences will start flocking to those deemed more trustworthy sources of news/analysis/thought leadership. Thus, the first mover advantage applies to these groups as well. Media figures are also taking the pledge, for example a conservative radio and podcast host, John Wells. Here is his statement for the PTP newsletter on taking the pledge:

> The lifeblood of my program to which my name is attached and therefore all who I call and who call me, friend, those who trust me to be honest with them, and most importantly in the Earthly realm, my family rely on truthfulness in what I do. And of supreme importance, God is watching. And listening. www.caravantomidnight.com

Similarly to the politicians, Wells' statement is designed to get him appropriate reputation rewards.

Liberal radio hosts are taking the pledge as well, for example Ethan Bearman, rated #57 talk show host in the country by Talkers magazine and frequent guest on CNN and Fox. His statement is as follows:

> Facts matter and the truth matters. With the state of communications allowing any bit of information, true or not, to instantly propagate across the globe, getting to the truth is as hard as its been in my memory. There are people who prey on others with falsehoods for monetary gain, political influence, and even pure malice. It is up to us to make sure the truth shines through the clouds of falsehoods. www.ethanbearman.com facebook.com/ethanbearmanshow twitter.com/ethanbearman Thank you! -Ethan Bearman

Like the two politicians, both Wells and Bearman know how to craft their statements to target an audience that cares about the truth, and get the appropriate reputation boost. Since both of

these talk show hosts announced their commitment to take the pledge on their programs, their listeners are now holding them accountable, along with PTP advocates who are assigned to this task.

Alternatives and Challenges

The current best alternatives to advancing truth in our political system focus on supporting the work of fact-checking organizations. Noble and worthwhile, these much-needed efforts unfortunately do not address the underlying problem of distrust in fact-checking organizations. For instance, according to a September 2016 Rasmussen Reports survey, only 29 percent of all likely voters in the US trust fact-checking of candidates' statements. The political disparity is enormous, and in-line with previous reporting on the partisan divide – 88 percent of Trump supporters do not trust fact-checkers, while 59 percent of Clinton supporters express trust for fact-checkers (Reports, 2016). This distrust for fact-checkers will not be solved by providing more fact-checking or faster, real-time fact-checking. Indeed, research shows that real-time fact-checking may actually make people more resistant to correct information (Garrett & Weeks, 2013). Such distrust can only be addressed by getting citizens to both care more about the truth and by providing credible information about who is truthful. The PTP aims to solve these problems through appealing to people's identities and getting them more emotionally invested into truth-oriented behavior, while also providing them with information about who are honest public figures. A secondary effect of the PTP may be to help legitimate trustworthy fact-checking organizations. Indeed, research suggests that training in media literacy is likely to reduce perceptions of bias by the media in reporting on controversial news stories, and the behaviors of the PTP are conducive to higher media literacy (Vraga, Tully, & Rojas, 2009).

Of course, the Pro-Truth Pledge may not work despite the problems with the current best alternatives. Virginity pledges have been shown consistently to delay the onset of sexual behavior (Martino, Elliott, Collins, Kanouse, & Berry, 2008). However, other research has shown that STD rates are comparable among those who took a virginity pledge and those who did not, potentially due to lower rates of condom use and testing by those took the pledge (Brückner & Bearman, 2005). Thus, the PTP may have mixed results in getting people to avoid sharing misinformation. Public figures may become afraid of signing on after a few suffered the reputational damage that comes from being listed as in contempt of the pledge. Likewise, politicians, media venues, and others who benefit from deceiving the voters will likely target the pledge as they see it gain ground. To fend off these attacks, the pledge organizers must work hard to reach across party lines to get diverse public figures from all sides of the political spectrum to commit to the

pledge, but this effort may or may not be successful. Another area of attack may be around the definition of misinformation as used by the PTP, for instance regarding potential bias in selecting fact-checking organizations. In part to ameliorate accusations of such bias, the PTP specifically decided to use the same fact-checking organizations as Facebook uses, since Facebook has a huge financial interest in using only the most high-quality fact-checking venues. Moreover, the PTP – unlike fact-checking organizations – only evaluates those who have chosen to sign the pledge; it is an opt-in mechanism, like the Better Business Bureau, as opposed to fact-checkers who fact-check statements that the fact-checking organization finds relevant.

Another finding that might be potentially problematic for the effectiveness of the pledge shows that citizens often use political figures they support as a guide to what they consider true or false, regardless of the facts (Swire et al., 2017). Counteracting this tendency requires that citizens develop trust and invest support into the Pro-Truth Pledge as a guiding mechanism for candidates they consider credible. Indeed, a number of people who have chosen to take the pledge have expressed that they would consider whether a candidate has taken the pledge a strong factor in choosing which candidates to support with their votes, money, and time.

Pro-Truth Pledge Impact: Case Studies

The PTP was launched in March 2017, and by March 11, 2018 had over 5800 pledge-takers. Of them over 500 are public figures, including such prominent names as Peter Singer, Steven Pinker, Michael Shermer, and Jonathan Haidt. Of these public figures, over 100 are public officials, such as Member of US Congress Beto O'Rourke. There are also over 50 organizations, such as Media Bias/Fact Check, The National Compass, Columbus Free Press, Fugitive Watch, and Earth Organization for Sustainability. Online and in-person groups dedicated to the PTP have emerged in over 20 US states, and are starting up in other states as well as abroad.

When asked for why they take the pledge, people generally report a desire to cast a vote against fake news and demonstrate a personal commitment to honesty. Some also discuss the desire to project a reputation as truth-tellers for the sake of gaining greater credibility among those with whom they engage.

We have performed some follow-up conversations with pledge-takers to determine whether the pledge impacted their behaviors. A private citizen, US Army veteran John Kirbow, stated how after taking the pledge, he felt "an open commitment to a certain attitude" to "think hard when I want to play an article or statistic which I'm not completely sold on." He found the pledge "really does seem to change one's habits," helping push him both to correct his own mistakes with an "attitude of humility and skepticism, and of honesty and moral sincerity," and also to

encourage "friends and peers to do so as well." Christian pastor and community leader Lorenzo Neal described how he "took the Pro-Truth Pledge because I expect our political leaders at every level of government to speak truth and not deliberately spread misinformation to the people they have been elected to serve. Having taken the pledge myself, I put forth the effort to continually gather information validating stories and headlines before sharing them on my social media outlets."

All others who chose to participate in follow-up conversations shared similar responses. It is important to note that follow-up conversations are limited by two factors: self-selection and self-reporting. After all, the people likely to respond are those who find the pledge beneficial and impactful. To address this concern, we also engaged in external observations of the behaviors of pledge-takers, and have observed instances where the pledge was involved with people retracting statements.

For instance, a candidate for Congress, Michael Smith, took the Pro-Truth Pledge (GoFundMe, 2017). He later posted on his Facebook wall a screenshot of a tweet by Donald Trump criticizing minority and disabled children. After being called out on it, he went and searched Trump's feed. He could not find the original tweet, and while Trump may have deleted that tweet, the candidate edited his own Facebook post to say that "Due to a Truth Pledge I have taken I have to say I have not been able to verify this post" (Imgur, 2017). He indicated that he would be more careful with future postings. In another case, Mark Kauffman, a photographer from New York, shared an article from OccupyDemocrats.com, a site shown by credible fact-checkers used by the PTP to be systematically unreliable. Other pledge-takers, following the behavior of asking people to stop using unreliable sources regardless of the credibility of the article, asked him to withdraw it, and he did so.

Pro-Truth Pledge Impact: Empirical Evaluation, First Study

Such case studies of interviews and observations, while illuminating, would be stronger if supported by more systematic quantitative data. Thus, we have conducted a survey evaluation of pledge-takers to see whether their sharing of information on social media was impacted by the pledge. We decided to target Facebook, as the most popular social media platform: 44 percent of US adults got news via Facebook in 2016 (Gottfried & Shearer, 2016). Our hypothesis was that taking the PTP will impact sharing on Facebook, both when people share news-relevant content themselves on their own Facebook profile, and when they engage in other venues on Facebook, such as Facebook groups or other people's profiles. Examples of engaging in other venues would include behaviors like asking people to retract incorrect statements, as was the

case with pledge-takers asking Michael Smith and Mark Kauffman to retract their statements. To test this hypothesis, we have conducted a longitudinal study of people who took the PTP and engage actively in sharing news-relevant content on Facebook. Since these people already care about the truth – otherwise, they would presumably not take the pledge - any difference between sharing behaviors before and after taking the pledge can be attributed to finding out about the pledge and taking it.

Method

We had participants fill out Likert scale (1-5) surveys self-reporting their Facebook engagement with news-relevant content on their own profiles and also with other people's posts and in groups before and after they took the pledge. The specific questions asked were in the form of the following:

> Before you heard about the Pro-Truth Pledge, to what extent did your behavior on your personal Facebook profile align with the Pro-Truth Pledge's 12 behaviors? Please give an estimate of 1 to 5, with 1 at lowest level of alignment to 5 being full alignment. Lowest alignment means sharing misinformation. Highest alignment means actively fighting lies and promoting truth. Remember that in this question, you are evaluating only your behavior on your personal profile, not your behavior in groups or in response to other people's posts.

All other questions followed a similar format. We asked a separate question about whether study participants wanted to clarify any aspects of the questions, and no one reported being confused by the questions.

To avoid the Hawthorne effect of study participants being impacted by observation, the study did not evaluate current behavior, but past behavior. We only recruited participants who took the pledge 4 or more weeks ago to fill out the survey, and asked them about their behavior after taking the pledge. Giving them this period also gave people an opportunity to have the immediate impact of taking the pledge fade from their mind, thus enabling an evaluation of the medium-term impact of the PTP on sharing news-relevant content. We recruited participants via Facebook posts and emails to people who took the pledge and were interested in receiving pledge updates soliciting participation in a study about the pledge, and participants were not given any incentives to participate. With these limitations, we were able to secure 24 participants. This study method was informed by the approaches used by studies of whether honor codes address cheating, which is the most comparable form of intervention to the PTP. Such studies similarly rely on self-reporting by students on whether they have cheated or not cheated (McCabe

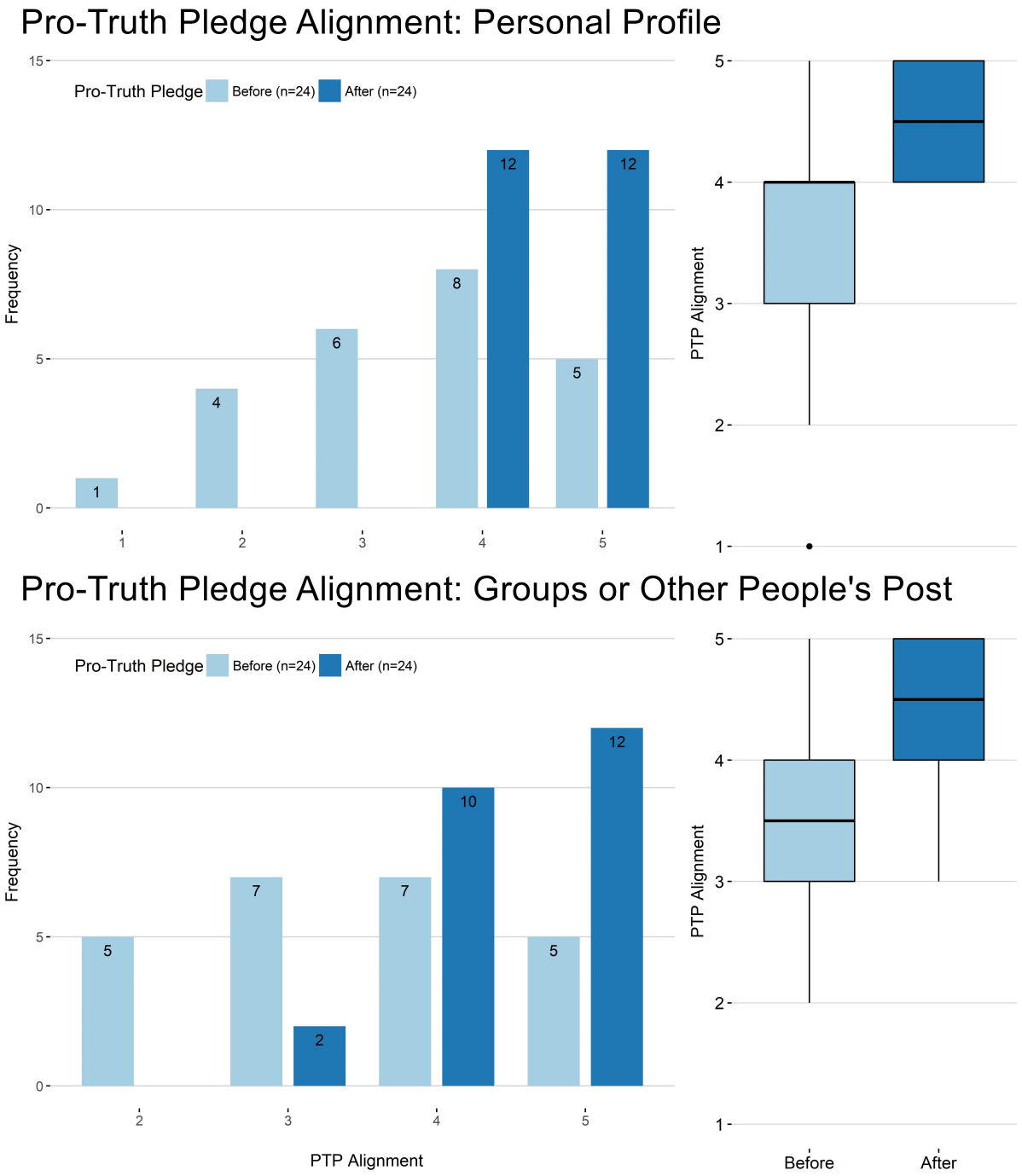
& Trevino, 1993; McCabe et al., 1999). Similarly, studies of whether virginity pledges delay sexual onset similarly reply on self-reporting (Brückner & Bearman, 2005). Thus, our method of evaluating the PTP faces the same problems faced by those studies: self-reporting and self-selection. Regarding the former, we cannot be certain whether, despite the clarity of the questions to the participants and the clarity of the behaviors outlined in the pledge, the study subjects gave accurate evaluations of themselves. Regarding the latter, there is a possibility that some people who failed to uphold the virginity pledge or the honor code might have chosen to avoid participating in studies of the impact of these interventions. However, given that these studies of honor codes and virginity pledges have been acknowledged as appropriate and influential in the literature despite the potential problems of self-reporting and self-selection, we have chosen to use a similar methodology to test a similar intervention.

Results

Our results show that taking the pledge results in a statistically significant increase in alignment with the behaviors of the pledge, both on one's own profile on Facebook and when interacting with other people's posts and in groups. Specifically, on one's own Facebook profile, the median alignment with the PTP score before taking the PTP is 4 (SD=1.14), and the median alignment score after taking the PTP is 4.5 (SD=0.51). We conducted an Asymptotic Wilcoxon-Pratt Signed-Rank Test to compare PTP alignment on one's own profile before and after taking the PTP. The null hypothesis for the test states that there is no significant score difference before and after taking the pledge and the alternative hypothesis proposes a significant difference. The results reveal a significant increase of PTP alignment after taking the pledge with a large effect size; $z = 6.12$, $p < 0.000$, $r = 0.88$. Based on the p-values, the null hypotheses can be rejected and the alternative hypothesis is accepted. These results suggest that taking the pledge really does influence alignment on one's own profile. The upper part of figure 2 represents the results visually.

For engaging with news-worthy content on other people's profiles, the median PTP alignment score before taking the Truth Pledge is 3.5 (SD=1.06). The median PTP alignment score after taking the Truth Pledge is 4.5 (SD=0.65). As before, we conducted an Asymptotic Wilcoxon-Pratt Signed-Rank Test to compare alignment in groups and other people's profiles before and after taking the Truth Pledge. The results show a significant increase of Truth Pledge Alignment after taking the pledge with a large effect size; $z = 6.11$, $p < 0.000$, $r = 0.88$. Again, the null hypothesis can be rejected and the alternative hypothesis is accepted.

Figure 1: Results of Study 1



PTP alignment is measured as: 1 = lowest alignment and 5 = highest alignment.

These results suggest that taking the Truth Pledge really does influence Truth Pledge Alignment in groups and on other people's posts. The lower part of figure 2 represents the results visually.

For sharing content on their own profile, 70.83% of participants (17 of 24 respondents) reported an increase of their PTP alignment after taking the PTP, eleven participants increased by one point on the alignment scale, five by two points, one by three points, while the rest maintained the same score. For sharing content in groups and on other people's walls, again, 70.83% of participants

reported an increase of their PTP alignment after taking the PTP, twelve participants increased by one point on the alignment scale, five by two, while the rest maintained their initial score. These results indicate that taking the PTP indeed significantly improves social media sharing, and thus the PTP is an effective intervention for addressing the scourge of fake news. The results also contradict the hypothesis that all those who take the PTP are already honest and that taking the pledge is simply a way to signal their honesty publicly. If that was the case, there would be no statistically significant increase in alignment with the truth-oriented behaviors in the PTP before and after taking the pledge.

Moreover, if we look at Figure 2 we see that about half of the survey participants were actually at 3 or below for the quality of their Facebook engagement before taking the pledge, and thus can hardly be called truth-oriented. It is only after taking the pledge that they improved their behavior. In fact, all study participants who scored at 3 or below reported some improvements in their behavior to align more with the pledge after taking it.

This study does not tell us whether the difference in sharing behaviors can be attributed simply to finding out about the pledge, or specifically to taking the pledge. Thus, the actual intervention we are measuring is the combination of "finding out about and taking the pledge" as opposed to the more fine-grained intervention of taking the pledge after finding out about it, or finding out about the pledge without taking it. While it may be an interesting question of whether just finding out about the pledge makes a difference in Facebook sharing behavior, or whether it's a matter both of finding out about the pledge and taking it, we had no reasonable way of reaching out and recruiting the many people who heard about the pledge without taking it. Moreover, from the perspective of social impact, it makes little difference: if finding out about the pledge and taking it is an effective intervention to decrease sharing misinformation, we can still ethically recommend encouraging people to learn about and take the pledge regardless of our lack of certainty about whether the actual mechanism involves just one or both components.

Pro-Truth Pledge Impact: Empirical Evaluation, Second Study

A weakness of the first study was its reliance on self-reporting. While a well-established academic method for studying behavior change, it has the weakness of being vulnerable to people's internal conceptions of themselves influencing their self-evaluations. After all, if someone took the Pro-Truth Pledge, they might be more likely to believe they were acting in accordance with the pledge – regardless of whether the PTP actually changed their behavior. Thus, while self-reporting is indicative of behavior change, we should not perceive it as conclusive evidence of the PTP's effectiveness. To evaluate more accurately whether the PTP actually results in

behavior change requires external observers to evaluate behavior change. Our second study involved enrolling some participants of the first study into a new study, where their behavior on Facebook was observed and coded by how well it aligned with the PTP. Thus, we could evaluate whether their self-reporting of behavior change correlated with actual behavior change. The second study included 21 people.

Method

Similarly to the first study, the second study avoided the Hawthorne effect of study participants being impacted by observation by evaluating past behavior. Study participants granted access to their Facebook profile to researchers, enabling researchers to take advantage of the Facebook Timeline feature to evaluate posts made by study participants after they took the pledge. Thus, the quality of sharing by study participants was not impacted by them knowing they were being observed, since they enrolled in the study after they already made the relevant Facebook posts that the researchers evaluated.

Researchers looked at the first ten Facebook posts with news-relevant content made four weeks after the pledge. The four week window enabled the initial impact of taking the pledge to fade from the minds of pledge-takers. Then, the researchers compared these ten posts to the first ten posts for the same period the year before the study participant took the pledge. The aim was to correct for any calendar-based differences in someone's Facebook sharing: for example, if the pledge-taker is a college student, they might do different types of sharing when they were taking classes vs. when they were on break. So if someone took the pledge on May 1, 2017, then the post-pledge sharing evaluation period began on May 29, 2017, and the evaluators looked at the first 10 posts made on and after that day. The pre-pledge sharing evaluation period began on May 29, 2016, and the evaluators looked at the first 10 news-relevant posts made on and after that day.

There were two coders who coded the posts of each of the 21 study participants, 10 before the participant took the pledge and 10 afterward, for a total of 420 individual pieces of data. The sharing was coded according to quality, from 1 of lowest level of alignment with the PTP, to 5 of highest alignment. Lowest alignment meant the content is misinformation, whether a news article or meme or personal post with news relevance. Highest alignment meant that the post actively promotes fighting lies and promoting truth. The coders evaluated both the post and the person's engagement in comments on that post as a total rating for each individual post. The point of making a total rating for each post - both the post and the person's engagement - is to approximate the impact of the person with each post on their social media followers, since

the followers will pay attention both to the original post, and the comments on the post. The guidelines were as follows:

> 1 (Lowest): the content is misinformation, whether a news article or meme or personal post.
>
> 2: the content is accurate, but comes from an unreliable source, even if the post itself does not contain misinformation.
>
> 3: the content is accurate, but it is satire without indicating it is satire; or the headline does not match the article without the person making the post indicating that the headline does not match the article; or it is a personal post or meme that does not cite sources.
>
> 4: the content is fine, with no problems
>
> 5 (highest): the content is specifically oriented to fighting misinformation and promote truth

Results

It is important to evaluate inter-coder reliability between the two coders, and a typical approach to this is to calculate Krippendorff's α, a measurement of agreement among coders of data, designed to indicate their reliability. It ranges from 0 to 1, whereas 0 indicates no agreement between the coders and 1 indicates perfect agreement. Strong coder reliability is indicated by $\alpha \geq .800$, and in cases where it is acceptable to have more tentative conclusions, $\alpha \geq .667$ is at the lowest acceptable limit (Krippendorff, 2004). The α for the two coders was .85, suggesting a good inter-coder reliability. We can assume that both of the coders agreed substantially on whether a Facebook post was in alignment with the Pro-Truth Pledge.

To evaluate the data, we took the average coding between the two coders, which left us with a single estimate per person per post (21 Truth-Pledge alignment scores before, and 21 scores after taking the Pro-Truth Pledge). Following this, a few checks were run to see whether the data meets the statistical assumptions for a paired t-test: 1.) normal distribution of the relevant variable and 2.) homogeneity of variance. To evaluate the first assumption, we conducted a Shapiro-Wilk test, and found that the data does not significantly differ from a normal distribution, W = 0.95, p-value = 0.08.

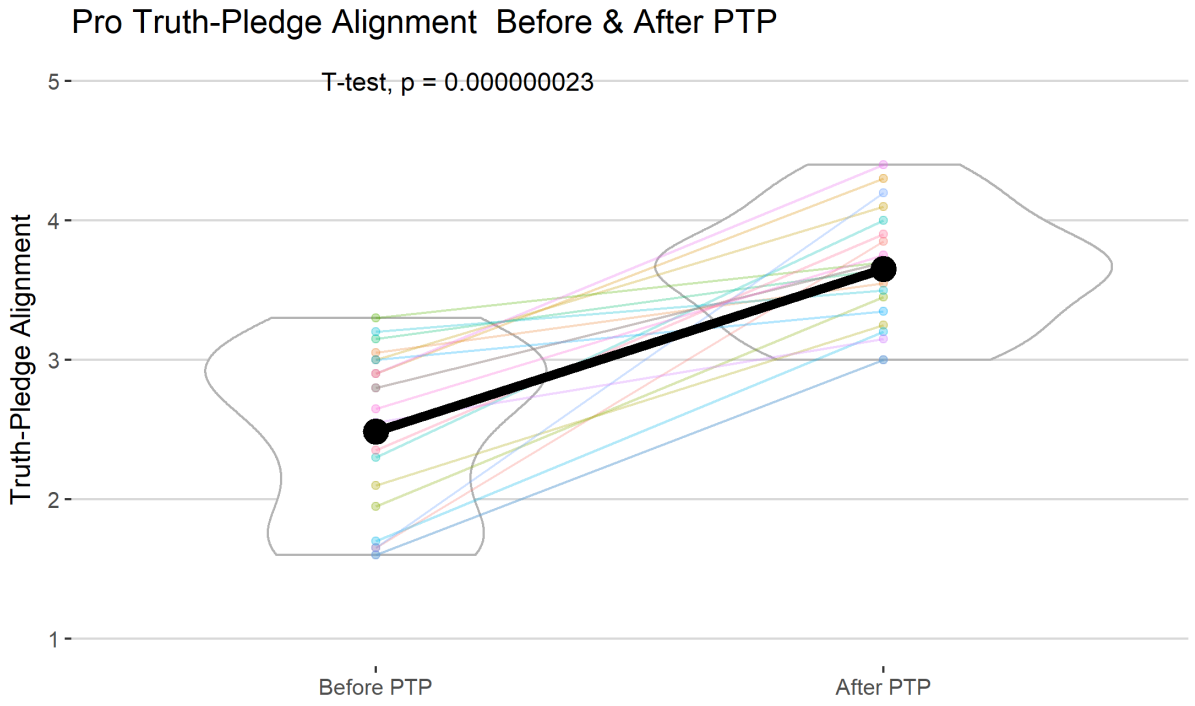
Following that, we conducted a Levene-Test to assess the homogeneity of variance. The test showed a p-value greater than 0.05, indicating that there is no significant difference in variances between the groups F = 4.069, p-value = 0.05. Thus, we can assume homogeneity of variance. Given that the statistical assumptions are met, a paired t-test can be estimated in order to examine whether Pro-Truth Pledge Alignment is significantly different after taking the PTP. The data can be seen in the table below:

Table 1: Descriptive Statistics of Data from the Second Study

	Time	N	Mean	SD
1	Before Truth-Pledge	21	2.49	0.60
2	After Truth-Pledge	21	3.65	0.41
3	Total	42	3.07	0.78

A paired-samples t-test was conducted to compare Pro Truth-Pledge Alignment before and after taking the Pro-Truth Pledge. There was a significant difference in the scores for Pledge Alignment *before* (M=2.49, SD=0.6) and *after* (M=3.65, SD=0.41) taking the PTP; $t(20) =$ -8.86, $p < 0.001$. An estimation of the effect size indicates that the found difference can be considered to be large (Cohen's d=-1.93). These results suggest that taking the PTP really does have an effect on inducing truthful sharing behavior on Facebook.

Figure 2: Results of Study 2



PTP alignment is measured as: 1 = lowest alignment and 5 = highest alignment.

The figure above provides a visualization of the results. The thick black line shows the median. The small colored lines represent change among individuals. Note that every individual who took the PTP has improved their sharing on Facebook to be more aligned with the PTP, some drastically.

Discussion

The first study showed that individuals self-report behaving more truthfully both on their own profiles and in other contexts on Facebook – such as on the profiles of friends and in groups – after taking the PTP. The improvement was large, with a clear statistical significance, of about 1 unit on a 1-5 scale. The second study focused on observing people's behavior on their own profiles, and confirmed that pledge-takers behaved more truthfully four weeks after taking the pledge. Again, the improvement had clear statistical significance, and was large, also about 1 unit on a 1-5 scale. In other words, people's self-reports about their improvement of behavior on their own profiles – and the extent of their improvement – were corroborated by external observers. We can thus assume that people also behaved more truthfully on other people's profiles and in groups, even though we have no realistic way of observing that. Overall, these two studies provide compelling evidence that people improved the honesty of their behavior on Facebook because they have heard about and signed the pledge, and there is no reason to believe they would have improved if they did not hear about and sign the pledge. **The combination of these two studies provides solid evidence that taking the PTP decreases the spread of misinformation on social media.**

By extension, the same finding implies that pledge-takers practice more truthful behavior in other areas of their civic engagement. Further research is needed to determine whether that is indeed the case. We also do not know whether presenting the PTP in a semi-voluntary context, such as when students are presented with an honor code with an implicit expectation that they sign it in order to attend the college of their choice, will maintain the impact of the PTP: further research is needed as well.

Conclusion

To solve the problem of private citizens sharing fake news and public figures engaging in deception to win and maintain power, we need techno-cognitive solutions, meaning ones that combine technology with psychological principles, according to prominent researchers in the field. The Pro-Truth Pledge, which combines psychology research with online mechanisms of implementation and propagation, and crowd-sources fact-checking, is one such intervention. It asks participants to commit to twelve behaviors, which are intended to counteract a number of

cognitive biases that contribute to people believing in and sharing misinformation, an essential aspect of the psychology research informing the content of the pledge itself.

In addition to committing to the behaviors of the pledge, pledge-takers are encouraged to share about taking the pledge on their social media, to put markers of taking the pledge on their online profiles, and to fact-check other pledge-takers, which is the crowd-sourcing component of the pledge. The PTP uses all four components shown by psychology research on environmental pollution as crucial to addressing tragedies of the commons (Milinsk, Semmann and Krambeck 2002, Van Vugt 2009). It provides information about the credibility of those who sign it, as well as information about what it means to orient toward the truth and what constitutes credible information sources. It appeals to the identity of people to desire to be honest and be perceived that way. Finally, it offers positive reputational rewards for honesty, taking advantage of the psychology research on incentives.

This techno-cognitive solution has shown some early signs of effectiveness. To be effective requires that: 1) we see evidence of people taking the pledge, and 2) we see evidence of people abiding by the behaviors of the pledge. We have clear evidence of people taking the pledge when they find out about it through venues like coverage by prominent media, through personal outreach, through word-of-mouth interactions on social media, and other means.

We also have growing evidence of its effectiveness in bringing about behavior change. Case studies indicate that at least some pledge-takers are moved to change their behavior after the pledge, including public figures. Two studies show that pledge-takers behave more truthfully on Facebook several weeks after taking the pledge. One of these studies relies on self-reporting, and another study relied on external observation of participants. In both cases, there was a large, statistically significant improvement in alignment with PTP behavior lasting for more than four weeks after subjects took the pledge. Further research needs to be done to test the effectiveness of this impact. In the meantime, this data suggests the pledge is valuable, and it is beneficial to encourage the widespread adoption of the Pro-Truth Pledge by all citizens and public figures who care about addressing the problem of fake news and post-truth politics.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *National Bureau of Economic Research*, *23089*. https://doi.org/doi:10.3386/w23089

Ariely, D., & Jones, S. (2012). *The honest truth about dishonesty: How we lie to everyone–Especially others* (Vol. 336). New York: Harper Collins.

Ariely, D., & Wertenbroch, K. (2002). Procrastination, deadlines, and performance: Self-control by precommitment. *Psychological Science*, *13*(3), 219–224. https://doi.org/doi:10.1111/1467-9280.00441

Barrera, O., Guriev, S., Henry, E., & Zhuravskaya, E. (2017). *Facts, alternative facts, and fact checking in times of post-truth politics (no. 12220)*. CEPR Discussion Papers.

Booth, R., Weaver, M., Hern, A., & Walker, S. (2017). Russia used hundreds of fake accounts to tweet about brexit, data shows. *the guardian*. Retrieved November 14, 2017, from https://www.theguardian.com/world/2017/nov/14/how-400-russia-run-fake-accounts-posted-bogus-brexit-tweets

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*, 617–620. https://doi.org/doi:10.1126/science.1183665

Brückner, H., & Bearman, P. (2005). After the promise: The std consequences of adolescent virginity pledges. *Journal of Adolescent Health*, *36*(4), 271–278. https://doi.org/doi:10.1016/j.jadohealth.2005.01.005

Burn, S. (1991). Social psychology and the stimulation of recycling behaviors: The block leader approach. *Journal of Applied Social Psychology*, *21*(8), 611–629. https://doi.org/doi:10.1111/j.1559-1816.1991.tb00539.x

Casadesus-Masanell, R., & Ricart, J. (2011). How to design a winning competitive position. *Harvard Business Review*, *89*, 100–107. https://doi.org/doi:10.1016/b978-0-7506-8548-1.50010-9

Cillizza, C. (2016). How the heck can voters think donald trump is more honest than hillary clinton? *Washington Post*. Retrieved from https://www.washingtonpost.com/news/the-fix/wp/2016/11/02/donald-trump-hasnt-told-the-truth-repeatedly-in-this-campaign-voters-still-think-he-is-more-honest-than-hillary-clinton/

Connelly, B., Certo, S., Ireland, R., & Reutzel, C. (2010). Signaling theory: A review and assessment. *Journal of Management*, *37*(1), 39–67. https://doi.org/doi:10.1177/0149206310388419

Correia, V., & Festinger, L. (2014). *Biased argumentation and critical thinking*. Bern: Peter Lang.

Dietz, T., Ostrom, E., & Stern, P. (2003). The struggle to govern the commons. *Science*, *302*(1907–1912). https://doi.org/doi:10.1126/science.1091015

Dunning, D. (2011). The dunning–Kruger effect. *Advances in Experimental Social Psychology*, *44*, 247–296.

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. In *Organizational behavior and human decision processes* (Vol. 105, pp. 98–121). https://doi.org/doi:10.1016/j.obhdp.2007.05.002

Farand, C. (2017). French social media awash with fake news stories from sources "exposed to russian influence" ahead of presidential election. *the independent*. Retrieved April 22, 2017, from http://www.independent.co.uk/news/world/europe/french-voters-deluge-fake-news-stories-facebook-twitter-russian-influence-days-before-election-a7696506.html

Fazio, L., Brashier, N., Payne, B., & Marsh, E. (2015). Knowledge does not protect.

Frijda, N., Manstead, A., & Bem, S. (2010). *Emotions and beliefs: How feelings influence thoughts*. Cambridge: Cambridge University Press.

Garrett, R., & Weeks, B. (2013). The promise and peril of real-time corrections to political misperceptions. In *Proceedings of the 2013 conference on computer supported cooperative work* (pp. 1047–1058). https://doi.org/doi:10.1145/2441776.2441895

Gino, F., Norton, M., & Ariely, D. (2010). The counterfeit self. *Psychological Science*, *21*(5), 712–720. https://doi.org/doi:10.1177/0956797610366545

GoFundMe. (2017). Fundraiser by michael smith: Michael w smith for congress 2018. Retrieved August 28, 2017, from https://www.gofundme.com/MichaelWSmith2018

Gottfried, J., & Shearer, E. (2016). News use across social media platforms 2016. *pew research center*. Retrieved May 26, 2016, from http://www.journalism.org/2016/05/26/news-use-acrosssocial-media-platforms-2016

Green, J., & Issenberg, S. (2016). Inside the trump bunker, with days to go. *Bloomberg Businessweek*, 2016–10.

Griskevicius, V., Tybur, J., & Bergh, B. (2010). Going green to be seen: Status, reputation, and conspicuous conservation. *Journal of Personality and Social Psychology*, *98*(3), 392–404. https://doi.org/doi:10.1037/a0017346

Guadagno, R., & Cialdini, R. (2010). Preference for consistency and social influence: A review of current research findings. *Social Influence*, *5*(3), 152–163. https://doi.org/doi:10.1080/15534510903332378

Guillory, J., & Geraci, L. (2013). Correcting erroneous inferences in memory: The role of source credibility. *Journal of Applied Research in Memory and Cognition*, *2*(4), 201–209.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Pantheon Books.

Hanley, N., & Folmer, H. (1998). *Game theory and the environment*. Edward Elgar.

Harding, G. (1968). Tragedy of the commons. *Science*, *162*(3859), 1243–1248.

Hatfield, E., Cacioppo, J., & Rapson, R. (1993). Emotional contagion. In *Current directions in psychological science* (Vol. 2, pp. 96–100). https://doi.org/doi:10.1111/1467-8721.ep10770953

Hopper, J., & Nielsen, J. (1991). Recycling as altruistic behavior. *Environment and Behavior*, *23*(2), 195–220. https://doi.org/doi:10.1177/0013916591232004

Imgur. (2017). Imgur: The most awesome images on the internet. Retrieved May 24, 2017, from http://imgur.com/a/A8IOY

Johnson, E., Shu, S., Dellaert, B., Fox, C., Goldstein, D., Häubl, G., & Larrick, R. (2012). Beyond nudges: Tools of a choice architecture. *Marketing Letters*, *23*(2), 487–504.

Jolls, C., Sunstein, C., & Thaler, R. (1998). A behavioral approach to law and economics. *Stanford Law Review*, *50*(5), 1471. https://doi.org/doi:10.2307/1229304

Katzev, R., & Pardini, A. (1987). The comparative effectiveness of reward and commitment approaches in motivating community recycling. *Journal of Environmental Systems*, *17*(2), 93–113. https://doi.org/doi:10.2190/xv00-dd4b-epeh-en5r

Kerin, R., Varadarajan, P., & Peterson, R. (1992). First-mover advantage: A synthesis, conceptual framework, and research propositions. *Journal of Marketing*, *56*(4), 33. https://doi.org/doi:10.2307/1251985

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology. thousand oaks.* California: Sage.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, *77*(6), 1121–1134. https://doi.org/doi:10.1037//0022-3514.77.6.1121

Kwong, J. (2017). "Why clinton lost: What russia did to. In *Control the american mind and put.*

Lord, K. (1994). Motivating recycling behavior: A quasiexperimental investigation of message and source strategies. *Psychology and Marketing*, *11*(4), 341–358. https://doi.org/doi:10.1002/mar.4220110404

Mann, H., Garcia-Rada, X., Houser, D., & D, A. (2014). Everybody else is doing it: Exploring social transmission of lying behavior. *PloS One*, *9*(10), 109591. https://doi.org/doi:10.1371/journal.pone.0109591

Martino, S., Elliott, M., Collins, R., Kanouse, D., & Berry, S. (2008). Virginity pledges among the willing: Delays in first intercourse and consistency of condom use. *Journal of Adolescent Health*, *43*(4), 341–348. https://doi.org/doi:10.1016/j.jadohealth.2008.02.018

Mazar, N., Amir, O., & Ariely, D. (2006). Dishonesty in everyday life and its policy implications. *Journal of Public Policy & Marketing*, *25*(1), 117–126. https://doi.org/doi:10.1509/jppm.25.1.117

Mazar, N., Amir, O., & Ariely, D. (2008). The dishonesty of honest people: A theory of self-concept maintenance. *Journal of Marketing Research*, *45*(6), 633–644. https://doi.org/doi:10.1509/jmkr.45.6.633

McCabe, D., & Trevino, L. (1993). Academic dishonesty: Honor codes and other contextual influences. *The Journal of Higher Education*, *64*(5), 522. https://doi.org/doi:10.2307/2959991

McCabe, D., Trevino, L., & Butterfield, K. (1999). Academic integrity in honor code and non-honor code environments: A qualitative investigation. *The Journal of Higher Education*, *70*(2), 211. https://doi.org/doi:10.2307/2649128

McDermott, R. (2004). The feeling of rationality: The meaning of neuroscientific advances for political science. *Perspectives on Politics*, *2*(4), 691–706.

Milinski, M., Semmann, D., & Krambeck, H.-J. (2002). Reputation helps solve the "tragedy of the commons". *Nature*, *415*(6870), 424–426. https://doi.org/doi:10.1038/415424a

Mullen, B., Brown, R., & Smith, C. (1992). Ingroup bias as a function of salience, relevance, and status: An integration. *European Journal of Social Psychology*, *22*(2), 103–122. https://doi.org/doi:10.1002/ejsp.2420220202

Myers, T., Maibach, E., Roser-Renouf, C., Akerlof, K., & Leiserowitz, A. (2013). The relationship between personal experience and belief in the reality of global warming. *Nature Climate Change*, *3*(4), 343–347. https://doi.org/doi:10.1038/nclimate1754

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*(2), 303–330.

Oskamp, S., Harrington, M., Edwards, T., Sherwood, D., M., S., & Swanson, D. (1991). Factors influencing household recycling behavior. *Environment and Behavior*, *23*(4), 494–519. https://doi.org/doi:10.1177/0013916591234005

Ostrom, E. (2015). *Governing the commons*. New York: Cambridge University Press.

Oxford Dictionaries. (2016). Word of the year 2016 is... *oxford dictionaries*. Retrieved August 28, 2017, from https://en.oxforddictionaries.com/word-of-the-year/word-of-the-year-2016.

Palmer, E. (2017). Spain catalonia: Did russian 'fake news' stir things up? BBC News. Retrieved November 18, 2017, from http://www.bbc.com/news/world-europe-41981539

Pennycook, G., Cheyne, J., Koehler, D., & Fugelsang, J. (2013). Belief bias during reasoning among religious believers and skeptics. *Psychonomic Bulletin & Review*, *20*(4), 806–811. https://doi.org/doi

Reports, R. (2016). The most comprehensive public opinion data anywhere (2016), "voters don't trust media fact-checking. Retrieved September 2016, from http://www.rasmussenreports.com/public_content/politics/general_politics/september_2016/voters_don_t_trust_media_fact_checking.

Schwarzer, R. (2008). Modeling health behavior change: How to predict and modify the adoption and maintenance of health behaviors. *Applied Psychology*, *57*(1), 1–29. https://doi.org/doi:10.1111/j.1464-0597.2007.00325.x

Selinger, E., & Whyte, K. (2011). Is there a right way to nudge? The practice and ethics of choice architecture. *Sociology Compass*, *5*(10), 923–935. https://doi.org/doi:10.1111/j.1751-9020.2011.00413.x

Sheldon, O., Dunning, D., & Ames, D. (2014). Emotionally unskilled, unaware, and uninterested in learning more: Reactions to feedback about deficits in emotional intelligence. *Journal of Applied Psychology*, *99*(1), 125–137. https://doi.org/doi:10.1037/a0034138

Shu, L., Mazar, N., Gino, F., Ariely, D., & Bazerman, M. (2012). Signing at the beginning makes ethics salient and decreases dishonest self-reports in comparison to signing at the end. In *Proceedings of the national academy of sciences* (Vol. 109, pp. 15197–15200). https://doi.org/doi:10.1073/pnas.1209746109

Shuster, S. (2017). Russia has launched a fake news war on europe. now germany is fighting back. *time*. Retrieved August 9, 2017, from http://time.com/4889471/germany-election-russia-fake-news-angela-merkel/

Silverman, C. (2016). This analysis shows how fake election news stories outperformed real news on facebook. *buzzfeed news*. Retrieved August 16, 2016, from https://www.buzzfeed.com/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Silverman, C., & Singer-Vine, J. (2016). Most americans who see fake news believe it, new survey says. *buzzfeed news*. Retrieved August 6, 2016, from https://www.buzzfeed.com/craigsilverman/fake-news-survey

Subramanian, S. (2017). Inside the macedonian fake-news complex. *wired*. Retrieved February 2017, from https://www.wired.com/2017/02/veles-macedonia-fake-news/

Sunstein, C., & Thaler, R. (2003a). "Libertarian paternalism is not an oxymoron. “university of chicago public law & legal theory working paper. https://doi.org/doi:10.2139/ssrn.405940

Sunstein, C., & Thaler, R. (2003b). Libertarian paternalism. *The American Economic Review*, *93*(2), 175–179. https://doi.org/doi:10.1017/cbo9780511790850.009

Sunstein, C., & Thaler, R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. New York: Penguin Books.

Sunstein, C., Thaler, R., & Balz, J. (2010). Choice architecture. *SSRN Electronic Journal*. https://doi.org/doi:10.2139/ssrn.1583509

Swift, A. (2016, September 14). Americans' trust in mass media sinks to new low. Gallup. Retrieved from http://www.gallup.com/poll/195542/americans-trust-mass-media-sinks-new-low.aspx.

Swire, B., Berinsky, A., Lewandowsky, S., & Ecker, U. (2017). Processing political misinformation: Comprehending the trump phenomenon. *Royal Society Open Science*, *4*(3). https://doi.org/doi: 10.1098/rsos.160802

The Annenberg Public Policy Center. (2017). Jamieson offers new name for fake news: “Viral deception” or vd. *the annenberg public policy center of the university of pennsylvania*. Retrieved from http://www.annenbergpublicpolicycenter.org/on-cnn-jamieson-offers-new-name-for-fake-news-viral-deception-or-v-d/.

Tyler, T., & Degoey, P. (1995). Collective restraint in social dilemmas: Procedural justice and social identification effects on support for authorities. *Journal of Personality and Social Psychology*, *69*(3), 482–497. https://doi.org/doi:10.1037//0022-3514.69.3.482

Van Lange, P., De Bruin, E., Otten, W., & Joireman, J. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology*, *73*(4), 733–746. https://doi.org/doi:10.1037//0022-3514.73.4.733

Van Vugt, M. (2001). Community identification moderating the impact of financial incentives in a natural social dilemma: Water conservation. *Personality and Social Psychology Bulletin*, *27*(11), 1440–1449. https://doi.org/doi:10.1177/01461672012711005

Van Vugt, M. (2009). Averting the tragedy of the commons. *Current Directions in Psychological Science*, *18*(3), 169–173. https://doi.org/doi:10.1111/j.1467-8721.2009.01630.x

Van Vugt, M., & Samuelson, C. (1999). The impact of personal metering in the management of a natural resource crisis: A social dilemma analysis. *Personality and Social Psychology Bulletin*, *25*(6), 735–750. https://doi.org/doi:10.1177/0146167299025006008

Verkuyten, M., & Nekuee, S. (1999). Ingroup bias: The effect of self-stereotyping, identification and group threat. *European Journal of Social Psychology*, *29*(23), 411–418. https://doi.org/doi:10.1002/(SICI)1099-0992(199903/05)29:2/3<411::AID-EJSP952>3.0.CO;2-8

Vining, J., & Ebreo, A. (1992). Predicting recycling behavior from global and specific environmental attitudes and changes in recycling opportunities1. *Journal of Applied Social Psychology*, *22*(20), 1580–1607. https://doi.org/doi:10.1111/j.1559-1816.1992.tb01758.x

Vogler, J. (2000). *The global commons: Environmental and technological governance* (2nd ed). Wiley.

Vraga, E., Tully, M., & Rojas, H. (2009). Media literacy training reduces perception of.

Westman, M., Eden, D., & Shirom, A. (1985). Job stress, cigarette smoking and cessation: The conditioning effects of peer support. *Social Science & Medicine*, *20*(6), 637–644. https://doi.org/doi:10.1016/0277-9536(85)90402-2

Zimmerman, R., & Connor, C. (1989). Health promotion in context: The effects of significant others on health behavior change. *Health Education Quarterly*, *16*(1), 57–75. https://doi.org/doi:10.1177/109019818901600108